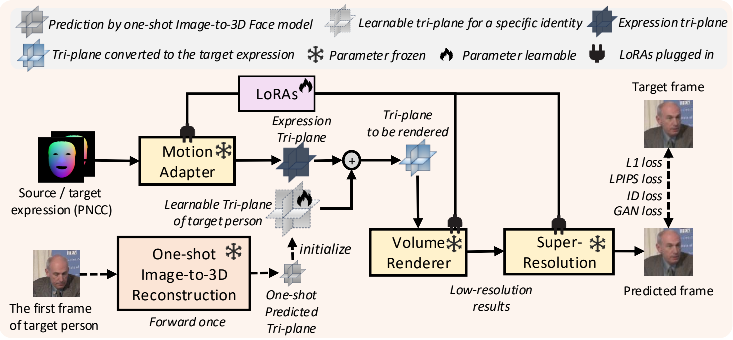
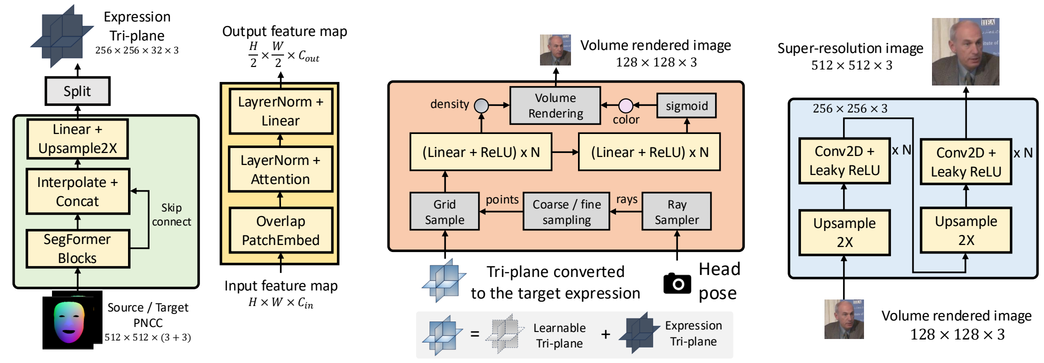

Prediction by one-shot Image-to-3D Face model
Learnable tri-plane for a specific identity
Expression tri-plane
Tri-plane converted to the target expression
Parameter frozen
Parameter learnable
LoRAs plugged in
LoRAs
Source / target expression (PNCC)
Motion Adapter
Expression Tri-plane
Tri-plane to be rendered
Target frame
Learnable Tri-plane of target person
L1 loss
LPIPS loss
ID loss
GAN loss
initialize
The first frame of target person
One-shot Image-to-3D Reconstruction
Forward once
One-shot Predicted Tri-plane
Volume Renderer
Super-Resolution
Low-resolution results
Predicted frame

Figure 1: New version of Figure 2 in the original paper. As suggested by Reviewer fWzj, we replace the cube with tri-plane to represent the 3D face NeRF field. And change the "3D face" with "tri-plane" in the label to improve clarity.



(a) SegFormer-based Motion Adapter

(b) SegFormer Block in Motion Adapter

(c) Volume Renderer

(d) Super-Resolution Module

Figure 2: Network details of each component in our generic model, as suggested by Reviewer fWzj.

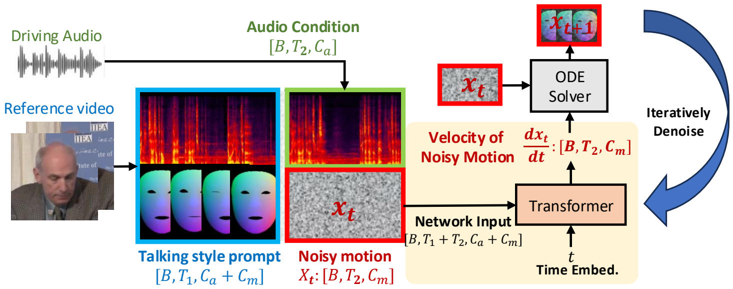


Figure 3: New version of Figure 3 in the original paper. As suggested by Reviewer fWzj, we improve the clarity of the inference process of ICS-A2M model.

Table 1: CMOS results on the lip-sync and expressiveness of various setting of our ICS-A2M model and the baseline ER-NeRF (ICCV 2023). CMOS score ranges from -3 to +3. Error bars are 95% confidence intervals.

Settings	CMOS-lip-sync↑	CMOS-expressive ↑
#1. Ours (MimicTalk with ICS-A2M model)	**0.000**	**0.000**
#2. Ours w.o flow matching	-0.438 ± 0.273	-0.895 ± 0.387
#4. Ours w. style vector	-0.361 ± 0.204	-0.532 ± 0.210
#5. Ours w. style encoder	-0.254 ± 0.188	-0.338 ± 0.225
#6. Ours w.o sync loss	-0.932 ± 0.349	-0.423 ± 0.296
#6. ER-NeRF (ICCV 2023)	-1.838 ± 0.486	-1.535 ± 0.422

Table 2: CMOS results on the style controllability and identity similarity of MimicTalk and StyleTalk. CMOS score ranges from -3 to +3. Error bars are 95% confidence intervals.

Methods	CMOS-style-control↑	CMOS-identitiy-similarity ↑
#1. MimicTalk (ours)	$\mathbf{0.549 \pm 0.225}$	$\mathbf{1.735 \pm 0.362}$
#2. StyleTalk (AAAI 2023)	0.000	0.000